On the impact of socio-economic factors on power load forecasting

Yufei Han, Xiaolan Sha, Etta Grover-Silva, Pietro Michiardi

On the Impact of Socio-economic Factors on Power Load Forecasting

Yufei Han
GridPocket

Xiaolan Sha
Eurecom

Etta Grover-Silva
GridPocket

Pietro Michiardi
Eurecom

Abstract

In this work, we study the importance of socio-economic factors of residential customers for estimating daily peak and total load, at fine temporal granularity. We do so by generating a compact set of heuristic rules using random forests. Compared with black-box time series prediction models in previous works, the rule set we obtain is highly interpretable and makes it easy to fuse human experts domain knowledge. In addition, we quantify and rank the importance of socio-economic features in the rule set for the forecast task. Our experimental analysis, which uses a publicly available dataset of over 3,800 households, providing consumption data for 1.5 year, highlights the superiority of tree-based models over state-of-the-art techniques that use support vector machines.

1. Introduction

Recent development of smart meter technologies have enabled energy providers to collect fine-grained (e.g., hourly) customer usage records. Analyzing large amounts of consumption data helps energy utilities to build user profiles, which opens opportunities to develop new customer oriented services. The liberalization of energy markets and added transparency of energy data motivates providers to offer energy tariff schemes better suited to customers' personal preferences: several energy providers have already put in place dynamic billing strategies and provide energy-saving services based on consumption data analytics [1], [2]. From the perspective of grid management systems, collecting and processing large amount of consumption records allow utilities to leverage energy production flexibly, to cope with peak usage, and monitor the health of the distribution grid at fine temporal granularities.

In this context, load forecast is a critical operation that involves the prediction of electricity consumption characteristics, e.g., demanded power level of grid substations or house-holds, over a variety of time horizons (minutely, hourly, daily or even yearly) [3], [4], [5], [6], [7]. Utilities greatly benefit from load forecasts for the managament of the energy supply-demand balance and the interaction of flexible load switching schemes [8]. In energy science, many studies focus on how to improve forecast accuracy using statistical time series processing algorithms [3], [6], [9], [10], [11], [17]. In contrast, little attention has been devoted to studying the impact of socio-economic factors in building customer usage profiles. Like any other private consumption records, e.g., online shopping histories, electricity usage is also a consumer oriented process, which can be inferred from customers' behavioral habits and householding characteristics [2], [12], [13]. As opposed to historical consumption associated with proprietary meters linked to a specific house, socio-economic factors are bound to customers and are expressed independently of a specific provider. The association between such factors and energy consumption patterns allows new incumbents to estimate energy demand profiles, even if few and incomplete historical consumption records are available. Moreover, although smart meter technology is emerging in the landscape of energy-related industry, energy providers still stand at the very early stage of deploying new generation meters. For most end-users, energy providers are unable to collect fine-grained historical consumption records.Therefore socio-economic customer profiles turn out to be a quintessential asset that should be exploited.

The aim of our work is two-fold: *i)* we propose a heuristic forecast model of customers' daily energy usage patterns based on their socio-economic factors and, *ii)* we identify the most important factors for forecasting purposes. As such, the results of our work can help utilities to investigate characteristics of energy usage of each end-user, avoiding verbose questionnaires and reducing the intrusiveness of private information collection. Our empirical study uses a

large-scale public dataset of electricity usage patterns of residential consumers. We proceed by defining two behavioral indicators, namely *peak load* and *total electric consumption* as the forecasting targets. Then, we design a forecasting model based on random forests [14], [15] to model the association between customers' socio-economic factors and the two indicators. Essentially, our ultimate goal is to estimate energy usage profiles based on their socio-economic attributes. Applications of the methods we study in this work are numerous. Based on the two selected indicators, energy providers can set a proper power limit for each customer and protect the grid from risk of overloading during daily peak hours [3], [10], [11]. Utilities can use the forecasting results to tune billing prices and energy production dynamically, according to variation of energy usage [4], [5]. For individual customers, the forecasting model can help them adjust their energy usage and reduce their bills [16].

The remainder of the paper is organized as follows. A review of the related research progress in load forecasting and a description of the necessary background on random forests is given in Section 2. In Section 3, a brief overview the statistical properties of the dataset used is performed along with a quantification of the information conveyed by the socio-economic factors extracted from the dataset. In addition, the feasibility of using data-driven models to estimate energy usage patterns through socio-economic factors is discussed. Section 4 describes the experimental configuration in detail. Moreover,the forecasting performance of this approach is compared to the current state-of-the-art. In doing so,the stability of our approach is investigated and a quantified ranking of the importance of the socio-economic features is defined for the forecast task.Section 5 concludes the paper.

2. Background and Related Work

In this Section, current efforts toward the design and operation of load forecast mechanisms is examined andthe necessary background on the techniques used in this paper is provided, namely decision trees and random forests.

Current Approaches to Load Forecasting. Load forecasting has received a lot of attention lately: most research work in this domain can be categorized into three categories.

Firstly, time-series based methods are used to model temporal causality of energy demand between the past and future [5], [6], [7], [9], [17], [18], [19], [20]. Popular machine learning algorithms, such as ARIMA [20], SVM [5], [19] and neural networks [17], [18], build black-box models of temporal dynamics. Thus, prediction is achieved following the learned functional mapping between past and future energy usage patterns. Accurate as they are, they offer no insights on the underlying phenomenon behind the time evolution of energy demand, and the predicitive models cannot be easily generalized.End user models are commonly userd as an alternative to black-box methods [1], [3], [4], [6], [8], [11], [16].This model requires information about housing conditions, electrical appliance usage and environmental factors. The key idea of the this model is to *disaggregate* daily energy consumption into elementary components – including heating/cooling, water usage, cooking and other behaviors – which are used to interpret the temporal variation of clients' energy demand. Thus, forecasting models use relations between the collected information and energy consumption profiles. The shortcoming of end-use models is that forecasting performance depends heavily on the quality of available information, which makes them sensitive to noise.

Econometric methods [3], [4], [6], [12], [13], [16] combine the above two techniques. Such models estimate the relationship between energy consumption profiles and the factors influencing consumption behavior, using the least fitting error criterion as in time-series based methods. Econometric models are built by learning the mapping from pairs of input factors and output energy consumption profiles automatically, which is appealing for realistic application deployments. Hence, this category has gained popularity recently.

Although energy demand-supply does not closely obey typical market laws, the purchase-consumption relation between utilities and end-users indicates energy usage is intrinsically a type of consumption behavior similar to on-line purchasing in e-commerce. The basic motivations driving clients to purchase more electricity is to satisfy their living requirements. Therefore, clients' comfort preferences, income levels, family structure, residential status, electrical appliances and environmental features, or, in short, their *socio-economic status*, can therefore change electricity usage behavior dramatically, which in turn determines consumption profiles.

As discussed above, end-use model explicitly integrate such factors in the forecasting model. However, they require human experts' interference to tail the

input factor set. To the best of our knowledge, the integration and ranking of socio-economic features for load forecasting through automatic learning from energy usage data has remained elusive in the literature. Using the methodology described in this work, we study how to appropriately, and automatically select socio-economic factors to build an energy forecast model. In doing so, we unveil which are the most relevant features that contribute to accurate predictions, and which can be safely dismissed as they contribute little to the predictive power of our model.

Random Forest Models.

The forecasting model in this work is based on random forests. Its concept dates back from the indenpendent work of Tin Kam Ho and Amit and Geman [14], [21] as a an ensemble variant of decision tree [14], [15]. The central idea is to construct a set of decision trees independently by bootstrapping training data. The output of a random forest is obtained, e.g., through majority voting or with a simple average of the output of each individual tree model. Each decision tree in a random forest is a tree-like rule chain based on input variables for classification or regression. Each rule is a branch-split operation comparing one input variable with a predefined threshold. An input sample is then forwarded to different decision branches for finer analysis. The hierarchical split-branch operations form a coarse-to-fine *white box model*, which can explain *explicitly* how input factors are combined to achieve a complete decision making procedure.

The main motivation of random forests stems from the fact that a single tree model is sensitive to noise and produces predictions with large variance. Random forests mitigate the issue by injecting randomness into the tree structures and finally aggregating the output of the decision trees built independently. First, bootstrap sampling helps produce different randomly sampled training data subsets. Then, a decision tree is trained using each subset. In addition, input attributes are also bootstrapped, and the selected attribute subsets are used to learn the tree structure: as such, individual decision trees are largely independent one from each other. Averaging the outputs over several learned trees can substantially reduce the variance and improve stability in the final decision output. Despite a seemingly complex procedure, random forests can be used to assess the importance for each input attribute.

The properties of random forest makes itself a perfect fit for the goal of our work. One further aspect that has not been addressed so far, pertain to the scalable implementation of a random forest algorithm. Although out of the scope of this paper – and thus excluded from our contributions – significant development was dedicated to *parallel design* of a random forest algorithm. We did so by relying on the *functional paradigm* offered by data-intensive scalable computing systems such as Apache Spark [22]. Essentially, the proposed algorithm follows the *MapReduce* programming model, and the forecasting models are built by executing on a parallel platform, particular attention was given to the iterative nature of the random forest process: intermediate data materialization was avoided (which contributed to a large extent to poor runtime performance) using in-memory, reliable data structures. We omit further details of our algorithm to avoid distracting readers from the goal of this work.

3. Load Forecasting using Socio-economic Factors

A detailed description of the dataset used in this work is provided, which identifies the features used to learn the proposed model. The statistical properties of such features are analyzed with emphasis on socio-economic indicators. As a general remark, the large volume of the available training samples offers a large coverage of energy and socio-economic factors, which in turns improve model generability.

3.1.The dataset

The CER ISSDA dataset is a publicly available energy consumption trace, collected by the Irish Commission for Energy Regulation (CER) in a smart meter study [23]: it contains electricity consumption data of 4,225 private households and 485 small / medium enterprises; the trace covers 1.5 years (from July 2009 to December 2010). For each customer, the daily load curve is sampled every 30 minutes: energy data can be thought of as a series of timestamps and energy readings. In addition to energy data, the dataset includes a series of survey sheets and answers for each consumer, describing their housing condition, occupancy, employment status, income level, social class, appliance usage information and other socio-economic factors. 41 survey questions belonging to five categories were carefully selected, that allowed to build consumer profiles based on heating and lighting behavior, hot water and other electrical appliances usage: a complete list of features is shown in Tab. 1. For this work, users with less than 10% of survey

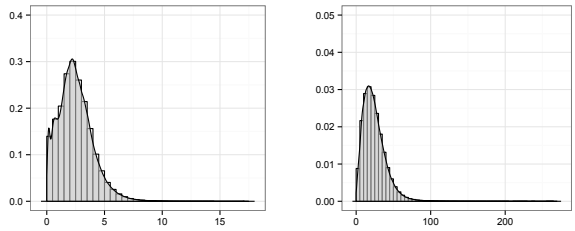
Category	Features
Household Profiles	Total number of occupants Number of occupants $>$ 15 years old Number of occupants $<$ 15 years old
Chief Income Owner	Gender Age Employment status Social class of the chief income Education level
Behavior Related	Interests in reducing bills by installing smart meters Interests in helping protect environments Interests in helping others who live with you to reduce bills Having succeeded in reducing the energy cost Having made changes in the energy consumption Willing to reduce energy consumption to protect the environment Having the knowledge to reduce the energy cost
House Characteristics	Type of the house The year when the house is built Number of bedrooms The proportion of double gazed windows in the house
Main Electrical Appliances	Type of the heating appliances in the house Type of the appliances to heat water in the house Type of the cooking appliances Type of the heating energy control system Type of the heating water control system Whether the house is always kept with adequate temperature
Other Electrical Appliances	Whether there is a washing machine Whether there is a tumble dryer Whether there is a dishwasher Whether there is an instant electric shower facility Whether there is an electric shower which pumps from hot water tank Whether there is an electric cooker Whether there is an electric heater Whether there is a stand alone freezer Whether there is a water pump or electric well pump or pressurized water system Whether there is an immersion
Weather	Heating Degree Days Cooling Degree Days Daily Average Humidity
Temporal Features	Is Holiday Month of the Year Day of the Week

Table 1: Breakdown of features extracted from surveys in the dataset and additional environmental factors.

coverage are eliminated, which results in a dataset of 3,822 households.

3.2Predictors used in the forecasting task

In addition to the socio-economic factors outlined in Sec. 3.1, we complement the predictors we use in the forecasting task with *environmental features*, which



(a) Daily Peak Load (b) Daily Consumption

Figure 1: Probability density function of the two forecasting targets.

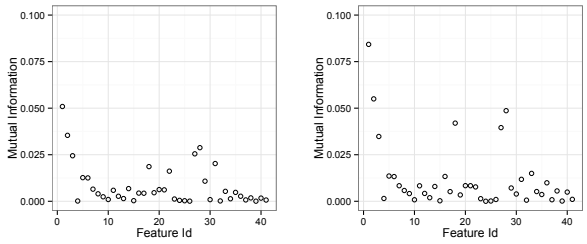
are significant for energy consumption profiling [18]. The first two are *heating degree days* and *cooling degree days*, evaluating quantitatively the needs to start heating or air conditioning appliances to keep an adequate environmental temperature. We also consider the *daily average humidity*, which represents the average air humidity level during a one-day interval: indeed, days with the same temperature but higher humidity level generally needs more energy to keep warm in winter or cool in summer. Finally, we include *month of year*, *day of week* and *holiday index*: the first feature represents the seasonal weather change, directly and strongly affecting energy usage profiles of all involved customers; the others features differentiate occupancy patterns of residential customers.

The input of our data-driven forecasting model includes a total of 41 factors, that are summarized in Tab. 1. Note that our predictors are of mixed formats, including both numeric and categorical types. As a consequence, a random forest model – which treats predictors uniformly in the splitting operation – is a more appropriate choice as compared to previously used statistical regressors, such as Support Vector Machine (SVM) [24], for which categorical predictors are problematic.

3.3Statistical profile of the load forecast task

The statistical profile of the dataset can be constructed using well-known feature ranking statistics to provide formal grounding to the key idea of this work: the goal is to show that socio-economic factors play an important role as predictors for a forecasting model.

As we observe in Fig. 1, the distributions of daily peak loads and daily total consumption are light-tailed:



(a) Daily Peak Loads (b) Daily Consumption

Figure 2: Mutual information of each feature in the prediction of peak load consumption and daily consumption.

the majority of the measurements locate around the mean, and there are no outliers.[1]

mutual information criterion [15], [25] was then used to illustrate the correlation between each input factor and the forecasting target: the larger the mutual information, the higher the correlation between the two random variables. Specifically, Kullbeck-Leiber divergence of the product of marginal distributions of two random variables x and y from the joint distribution of them was used, as illustrated in Eq. 1:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)\, p(y)} \right) \quad (1)$$

Fig. 2a and 2b report the mutual information for all the 41 socio-economic factors extracted from our dataset. More specifically, Tab. 2 reports the top-15 factors (sorted in descending order of mutual information), that are mostly related to daily peak loads and daily total energy usage respectively. We emphasize that here, the mutual information is not used to rank predictors for the purpose of feature selection.

From Tab. 2, occupancy attributes, the number of bedrooms, age of the clients, employment status of the clients, usage of tumble dryer and usage of dishwashers, are the common key elements influencing both peak loads and total amount of energy consumption, which are related to clients consuming capacities, potential electricity needs to keep proper temperatures and daily housework.

1. In general, outliers are often caused by either mistakes in measurement or faulty meters, and they usually increase the bias of data-driven models, leading to performance degradation.

Prediction of daily peak loads	**Prediction of daily total energy usage**
Total number of occupants	Total number of occupants
Number of occupants $>$ 15 years old	Number of occupants $>$ 15 years old
Whether there is a dishwasher	Whether there is a dishwasher
Whether there is a tumble dryer	Number of bedrooms
Number of occupants $<$ 15 years old	Whether there is a tumble dryer
Whether there is an electric cooker	Number of occupants $<$ 15 years old
Number of bedrooms	Whether there is a stand-by freezer
Type of the cooking appliances	Age
Age	Type of the house
Employment status	Employment status
Whether there is an instant electric shower facility	Whether there is an electric cooker
Willing to reduce energy consumption to protect the environment	Month of year
Social class of the chief income	Type of the heating appliances in the house
Type of the heating appliances in the house	Type of the appliances to heat water in the house
Type of the appliances to heat water in the house	Social class of the chief income

Table 2: Top 15 factors in prediction tasks using mutual information criterion

For the purpose of the forecasting task, the analysis indicates that cooking appliances – because they consume a lot of energy – largely contribute to the magnitude of the daily peak energy consumption. Instead, such appliances only marginally affect the the daily total energy consumption, as their usage is restricted to a small interval of time. In contrast, our analysis indicates that house category and usage of a standalone freezer device are key factors for the daily total energy consumption, but are weakly related to the daily peak load. Such results follow common sense: the type of the house is related to the thermal performance and inertia of the house, which crucially affects the total amount of energy – required to maintain a desirable temperature in the house. It has been shown that a data-driven model based on socio-economic factors may indeed find associations between user profiles and their consumption behavior. This method will be compared with the state of the art for validation.

4. Experimental Evaluation

The performance of the proposed random forest model is evaluated for forecasting daily peak loads and total electricity consumption of each residential user. In addition, the proposed method is compared to a state-

of-the-art approach of time series forecasting, Support Vector Machine (SVM), for validation purposes. We conclude with an analysis of the importance ranking of input socio-economic and environmental factors.

4.1. Comparison of Forecasting Models

The experimental methodology is as follows. For each of the 3,822 users and for the 1.5 year duration of our dataset, daily peak load measurement (in KiloWatt) and daily total electricity consumption (in KiloWattHour) are extracted to build the prediction targets. Input predictors are the 41 features described in Sec. 3. Therefore, we obtain 1,615,541 training samples, containing the input-output pairs for each of the forecasting tasks.

Metrics. For all models involved in our comparative analysis, we measure the forecasting performance – precisely, the accuracy of model fiiting – using *coefficient of determination* R^2 [15], [25]: it is defined by the ratio between the sum of square regression residual error and total sum of squares of the forecasting target. Large values of R^2 indicate an agreement between the model and the underlying real output, which translates in higher accuracy. Note that for really biased predictions, the coefficient can be zero or negative, while it is upper bounded by 1.

In this work, the relative ratio coefficient is used instead of the sum of square error, such that the variance of underlying forecasting target is taken into consideration to define an accuracy indicator. Higher variance implies that it is more difficult to construct a model to explain the overall data distribution. Thus, the R^2 coefficient is a more fair evaluation criterion of the model forecasting power than the sum of square errors. Eq.2 formalizes the performance metric, where y_i and f_i are the ground truth and the corresponding predicted value and $\bar{y}$ is the empirical expectation of the ground truth.

$$R^2 = 1 - \frac{\sum_i (f_i - y_i)^2}{\sum_i (y_i - \bar{y})} \tag{2}$$

Forecasting models. The main parameter of random forests is the number of trees constructed in the ensemble model to perform a final vote. In the experiment, this parameter is varied in the range $\{50, 100, 200, 300, 400\}$, and the stability of forecasting performance is investigated. Once the random forest model is constructed, the out-of-bag error [15], [25] of the random forest is used directly as the estimation of the model generalization error. Note that the split-branch operation in the random forest constructs a piecewise linear regression model, which can handle non-linearly distributed data without introducing additional algorithmic components.

Model	Peak Load: R^2	Total Load: R^2
RF(50)	0.5214	0.7151
RF(100)	0.5240	0.7261
RF(200)	0.5127	0.7236
RF(300)	0.5208	0.7100
RF(400)	0.5157	0.7070
SVM	0.4816	0.6572

Table 3: Generalization error of peak load and total electricity consumption prediction, with the Random Forest Model (RF) and SVM. Results for the RF model include values for the tree parameter (in parenthesis) we used.

Since the training set consists in more than 1.8 million records, a standard approach such as SVM can not afford the construction of a huge kernel matrix to enforce a non-linear regression *in memory*: hence, we use a linear kernel SVM. The training configuration of SVM is carefully selected through a 5-fold cross-validation. After fixing the configuration parameter, the SVM training ends by providing a linear regressor: 5-fold cross-validation is employed again to estimate its generalization error. Overall, the computational complexity of building the random forest model is $O(mn\log(n))$, where m is the number of trees, n is the number of bootstrapped training samples to construct each tree in the model. Considering that m is usually much smaller than n, the complexity of random forest is comparable to or smaller than the quadratic programming cost of SVM [15], [24], which is between $O(n^2)$ and $O(n^3)$. It is noted that the implementation of the random forest uses the MapReduce programming model, which is compiled to be executed on the Apache Spark platform. In the interest of space, the algorithmic details of the approach is not provided. However its capability goes beyond the embarassingly parallel nature of the random forest model, by also using parallel algorithms to build individual trees. As such, the method can *scale to large datasets*.

Results. Tab. 3 summarizes the generalization error for the propsed method and SVM. In general, this method outperforms SVM: precisely, the highest accuracy is achieved with an ensemble of 100 trees, for both forecasting tasks. Indeed, although more trees can reduce variance, they might increase the model bias.

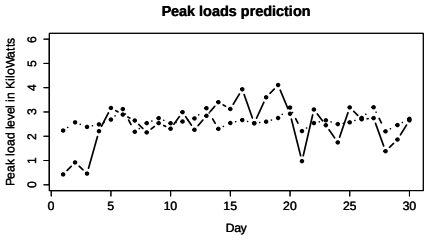


Figure 3: Prediction of peak load level of one user during one month.

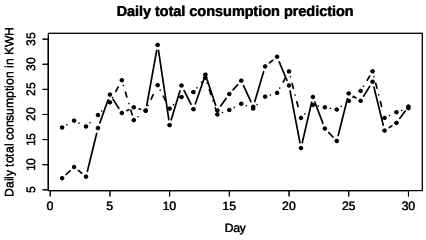


Figure 4: Prediction of total consumption usage of one user during one month.

There are two main reasons for the superiority of the approach. First, SVM treats categorical inputs as numerical variables, which hinders the task of defining an appropriate similarity measure: therefore, categorical information is essentially lost during the construction of the SVM model. Furthermore, due to the computational complexity and memory requirements, the SVM model is linear: thus, it cannot model the non-linear relation between socio-economic features and energy consumption profiles. In contrast, the random forest method does not suffer from such limitations. For illustrative purposes, Fig. 3 and 4 show two instances of the forecasting task for daily peak load and daily total consumption respectively: the solid line corresponds to the ground truth of daily peak load or daily total consumption measurements, while the dashed line represents the estimated values using the random forest model. These figures reveal that our method achieves higher predictive power for the total daily consumption than for the daily peak load, which confirms our findings in Tab. 3. The random forest model obtains distinctively higher R^2 in daily total consumption prediction. The results are explained as follows. Peak loads, as an instant power level measurement, are easily affected by householders' occupancy duration, appliances' working status, and in general, time-of-day related behavior. Such effects inject randomness into daily peak load patterns, which in turn are hardly represented comprehensively by the questionnaire used to derive input features of our model. In contrast, the daily total consumption, as a cumulated sum of energy consumption measurements, is less prone to random fluctuations due to time-of-day behavior. As a consequence, the model built with the approach contains a strong association to the static socio-economic features available in the dataset.

Prediction of daily peak loads	**Prediction of daily total energy usage**
Total number of occupants	Total number of occupants
Number of occupants $>$ 15 years old	Number of occupants $>$ 15 years old
Type of the cooking appliances	Average daily heating degree days
Whether there is an electric cooker	Number of bedrooms
Average daily heating degree days	Whether there is a dishwasher
Whether there is a tumble dryer	Month of year
Whether there is a dishwasher	Whether there is a tumble dryer
Type of the appliances to heat water in the house	Whether there is an electric heater
Whether there is an instant electric shower facility	Whether there is a stand-by freezer
Whether there is an electric heater	Type of the appliances to heat water in the house
Age	Whether there is an electric cooker
Whether there is a stand-by freezer	The year when the house is built
Education level	Whether there is an electric shower which pumps from hot water tank
Having made changes in the energy consumption	Employment status
Number of bedrooms	Age

Table 4: Top 15 factors in prediction tasks.

4.2Importance Ranking of the Socio-economic and Environmental Factors

An important by-product of the proposed random forest model is the inherent ranking of each input factor toward the forecasting task. The quantitative scores produced by our method are used to evaluate the importance of association between the socio-economic factors and daily energy usage profiles.

The ordering rank of input factors were recorded, which were derived through bootstrap sampling the training data during the construction of the random for-

est model. Therefore, the mean and standard deviation of the feature importance score of each input factor can be estimated. Note the standard deviations are at least three times smaller than the mean values for all top 15 factors. It indicates the ranking result is statistically stable with respect to the random forest parameter, *i.e.,* the number of trees used to build the model and training data sampling bias. Hence the factor importance ranking is beyond only data-driven output, but represents underlying association rules contained in the data. Thus the model built with 400 trees was used as a reference in the following discussion.

The experimental results are summarized in Tab. 4, which reports the top-15 input factors for predicting peak loads and total energy consumption. The ranking obtained is consistent to that obtained using the mutual information criterion in Sec. 3. The number of occupants is ranked as the most important feature, affecting both daily peak load level and the total amount of energy consumption, which follows the intuition of a direct relation between household size and energy consumption. General appliances and age are common features for the two forecasting tasks, but they are ranked higher for the daily peak load than for total consumption prediction task. Indeed, appliances contribute to a large extent to instantaneous peaks in energy consumption, and are generally responsible for aggregate energy consumption as well. Appliances that have a daily cycle (e.g. freezers) and average daily heating degree days are also shared features in two forecasting tasks. However, they are ranked much higher for predicting the total daily energy consumption, as they contribute to a steady energy consumption, rather than representative of peak demand. Some important distinctive traits between the ranking obtained by the proposed method and that based on mutual information are examined. For peak load forecasting, the random forest method boosts the importance of the water heating appliances, while it downgrades indirect factors, such as the number of bedrooms, and completely neglect some economic factors, like social class of the chief income and employment status. For total daily energy consumption, the random forests method assigns higher importance scores to factors related to heating usage – including environmental heating degree index, month of year, the year when the house is built, the number of bedrooms – while it decreases the importance of social class, age and employment status of the owner. Overall, the ranking produced by the random forest method is in line to that obtained through a manual inspection of the data by domain experts. It should be noted that environmental temperature factor does not play an important role in both forecasting tasks. The possible reason is high heating loads at all times due to Irish weather status, which limits the capability to trend the consumption data with varying temperatures. Furthermore, the collected temperature data is lack of precision for specific region, coupled with lack of exact knowledge of the geographic location of each household.

5. Conclusion

In this paper, the issue of forecasting daily electricity usage patterns and evaluating the influence of clients' socio-economic factors on energy consumption has been addressed.

A forecasting method based on random forest that seamlessly incorporates both clients' socio-economic factors and environmental factors has been proposed, resulting in an ensemble of split-branch decision rule chains, whereby a voting mechanism is used to achieve a stable prediction. The rule-chain based structure enabled the explaination of how each input features contributes to energy consumption forecasting, thus unveiling the underlying physical association between predictors and the daily energy usage pattern. These learned associations can either be used to discover unknown factors of energy consumption behaviors, or they can be applied as a complementary decision support to human experts.

Experimental results based on large-scale energy consumption records showed that the proposed random forest method dramatically outperforms a state-of-the-art SVM-based method, both in terms of prediction accuracy and in terms of scalability. Additionally, our work evaluated automatically importance of socio-economic factors in the forecasting tasks. The result can guide the construction of socio-economic surveys without requiring human intervention, leading to a decreased intrusiveness of measurement campaigns.

As shown in these experiments, daily peak load forecasting is a difficult task, compared to aggregate energy consumption. Future work intends to build detailed usage patterns of large power appliances that play a key role in determining peak loads and integrate historical peak load information to extend the input set of the forecasting model. In addition, we will provide details on our parallel algorithm design of the random forest model, and focus on its scalability properties by conducting a large-scale experimental campaign.

References

[1] Wood,G.,and Newborough,M.,"Dynamic energy-consumption indicators for domestic appliances: environment, behaviour and design", *Energy and Buildings*,vol.35,pp.821-841,2003.

[2] McLoughlin,F.,Duffy,A.and Conlon,M.,"Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables: An Irish case study",*Energy and Buildings*,vol.48,pp.240-248,2012.

[3] Tennesee Valley Authority, "Energy Vision 2020: Integrated Resource Plan/Environmental Impact Statement",Chapter 6, December,1995.

[4] Kolter,Z.,and Ferreira Jr, J.,"A large-scale study on predicting and contextualizing building energy usage", *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, pp.330-338, 2011.

[5] M.Centra, "Hourly Electricity Load Forecasting: An Empirical Application to the Italian Railways", *World Academy of Science, Engineering and Technology*,vol.5,pp.888-895,2011.

[6] Hesham K.A.,and Nazeeruddin, M.,"Electric load forecasting: literature survey and classification of methods", *International Journal of Systems Science*,vol.33,pp.23-34,2002.

[7] Humeau, S.F.R.J.,Wijaya, T. K., Vasirani, M. and Aberer, K.,"Electricity Load Forecasting for Residential Customers: Exploiting Aggregation and Correlation between Households",*Sustainable Internet and ICT for Sustainability (SustainIT)*,pp.1-6,2013.

[8] Aman, S.,Simmhan, Y.and Prasanna,V.K.,"Improving Energy Use Forecast for Campus Micro-grids using Indirect Indicators", *International Workshop on Domain Driven Data Mining (DDDM)*,pp.1-9,2011.

[9] Aung,Z.,Touky,M.,Williams,J.R.,and Herrero,S.,"Towards Accurate Electricity Load Forecasting in Smart Grids", *Proceedings of The Fourth International Conference on Advances in Databases, Knowledge, and Data Applications*, pp.52-57, 2012.

[10] Miranda,V.,and Monteiro,C.,"Fuzzy Inference Applied to Spatial Load Forecasting", *Proceedings of International Conference on Electric Power Engineering, 1999 (PowerTech Budapest 99)*, pp.35-40, 1999.

[11] Fu,C.W.,and Nguyen,T.T, "Models for long-term energy forecasting", *Proceedings of IEEE Power Engineering Society General Meeting*, pp.235-239,2003.

[12] Beckel,C.,Sadamori,Leyna.,and Santini,S.,"Towards automatic classification of private households using electricity consumption data", *Proceedings of the Fourth ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*, pp.169-176, 2012.

[13] Beckel,C.,Sadamori,Leyna.,and Santini,S.,"Automatic socio-economic classification of households using electricity consumption data",*Proceedings of the fourth international conference on Future energy systems*,pp.75-86,2013.

[14] Eiber,F.,"Pruning decision trees and lists",*PhD Thesis of The University of Waikato*,2000.

[15] Bishop,C.M.,"Pattern Recognition and Machine Learning",*Spring Verlag*,2006.

[16] Abreu, J.,and Pereira, F.,"Household Electricity Consumption Routines and Tailored Feedback", *Proceedings of ACEEE Summer Study on Energy Efficiency*,pp.193-206,2012.

[17] Kermanshahi B. S., and Iwamiya H.,"Up to year 2020 load forecasting using neural nets", *Electric Power System Research*, *Elsevier*,vol.24,pp.787-797,2002.

[18] Taylor,J.W.,and Buizza,R.,"Neural Network Load Forecasting With Weather Ensemble Prediction", *IEEE Transactions on Power Systems*,vol.17,pp.626-632,2002.

[19] Mohandes,M.,"Support vector machines for shortterm electrical load forecasting",*International Journal of Energy Research*,vol.26,pp.335-345,2002.

[20] El Desouky,A.A.,and Elkateb,M.M.,"Hybrid adaptive techniques for electric-load forecast using ANN and ARIMA",*IEE proceedings of Generation,Transmission and Distribution*,vol.147,pp.213-217,2000.

[21] Breiman,L.,"Random Forests", *Machine Learning*,vol.45,pp.5-32,2001.

[22] Apache Spark, http://spark.apache.org/

[23] "Eletricity Customer Behaviour Trial",*Commission for Energy Regulation*,Ireland,2011.

[24] Vapnik,V.N.,"The Nature of Statistical Learning Theory", *Springer Verlag*,*New York*,2002.

[25] Witten,I.,Frank,E.,and Hall,M.,"Data Mining: Practical Machine Learning Tools and Techniques",*Morgan Kaufmann*,2011.

[26] Fan,R.E., Chang,K.W., Hsieh,C.J., Wang,X.R., and Lin,C.J.,"LIBLINEAR: A library for large linear classification Journal of Machine Learning Research",*Journal of Machine Learning Research*,vol.9,pp.1871-1874,2008.